Responsible NLP Checklist

Paper title: *SciEvent: Benchmarking Multi-domain Scientific Event Extraction*
Authors: *Bofu Dong, Pritesh Shah, Sumedh Sonawane, Tiyasha Banerjee, Erin Brady, Xinya Du, Ming Jiang*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

□ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ A. Questions mandatory for all submissions.

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☑ A2. Did you discuss any potential risks of your work?
Limitations

☑ B. Did you use or create scientific artifacts? (e.g. code, datasets, models)

☑ B1. Did you cite the creators of artifacts you used?
3. SciEvent Benchmark, 4. Task Definition and 5. Experiment Settings

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The model and implementations in this study are publicly accessible. For our benchmark data (scientific abstracts), we will discuss the attributions fully in our public GitHub repository when we release the data. Reporting results for experiments does not violate any of the license of the used data.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3. SciEvent Benchmark, 4. Task Definition and 5. Experiment Settings

☒ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
Our data source is from publically available publications and filtered any Personally Identifying Info or Offensive Content.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We did not discuss these baselines in terms of their documentation, as we already cited their publically available publication and codebase.

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice.

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
3. SciEvent Benchmark

☑ **C. Did you run computational experiments?**

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5. Experiment Settings

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4. Task Definition and 5. Experiment Settings

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Descriptive statistics is not supportive to our experiments

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
4. Task Definition

☑ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix B codebook details

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Ethical Considerations and 3. SciEvent Benchmark

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
Our paper only use public scientific publication as data, and annotators are not giving any personal information. See B2 above, we will release full attribution when we release any public data.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The data is not related to any problem that needs approval from an ethics review board.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3. SciEvent Benchmark

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)